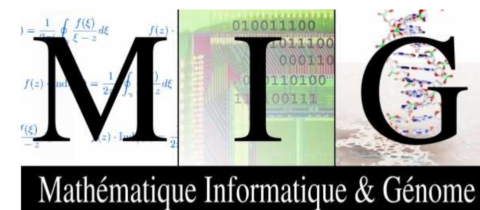
Les chaînes de Markov cachées : présentation et usage en analyse de séquences biologiques

Sophie Schbath

Unité Mathématique, Informatique & Génome

INRA – Jouy-en-Josas





Introduction

Pourquoi un modèle aléatoire ?

L'utilisation de modèles probabilistes pour l'analyse de séquences biologiques intervient dans de nombreux problèmes :

Pourquoi un modèle aléatoire ?

L'utilisation de modèles probabilistes pour l'analyse de séquences biologiques intervient dans de nombreux problèmes :

- est-ce qu'un événement observé est significatif ou simplement le fruit du hasard ?
  - fréquence ou présence d'un motif,
  - score d'alignement de séquences,
  - nombre de répétitions, etc.

Pourquoi un modèle aléatoire ?

L'utilisation de modèles probabilistes pour l'analyse de séquences biologiques intervient dans de nombreux problèmes :

- est-ce qu'un événement observé est significatif ou simplement le fruit du hasard ?
  - fréquence ou présence d'un motif,
  - score d'alignement de séquences,
  - nombre de répétitions, etc.
- modéliser l'alternance d'états dans une séquence et caractériser cette structure le mieux possible sur une séquence observée :
  - codant/non codant (introns/exons/intergénique),
  - transferts horizontaux chez les bactéries,
  - régions variables/constantes des virus, etc.

Pourquoi un modèle aléatoire ?

L'utilisation de modèles probabilistes pour l'analyse de séquences biologiques intervient dans de nombreux problèmes :

- est-ce qu'un événement observé est significatif ou simplement le fruit du hasard ?
  - fréquence ou présence d'un motif,
  - score d'alignement de séquences,
  - nombre de répétitions, etc.
- modéliser l'alternance d'états dans une séquence et caractériser cette structure le mieux possible sur une séquence observée :
  - codant/non codant (introns/exons/intergénique),
  - transferts horizontaux chez les bactéries,
  - régions variables/constantes des virus, etc.
- l'analyse de l'évolution des séquences au cours du temps, etc.

Modèles de séquences classiques

Soit $X_1, X_2, \ldots, X_n$ une suite aléatoire de lettres $X_i \in \mathcal{A}$.

Modèles de séquences classiques

Soit $X_1, X_2, \ldots, X_n$ une suite aléatoire de lettres $X_i \in \mathcal{A}$.

- Modèle de Bernoulli : M0
  Les X_i sont indépendantes et générées avec les probabilités

$$\mu(a) = \mathbb{P}(X_i = a), \quad \forall a \in \mathcal{A}$$

  Peut s'ajuster sur la fréquence observée des lettres d'une séquence.

Modèles de séquences classiques

Soit $X_1, X_2, \ldots, X_n$ une suite aléatoire de lettres $X_i \in \mathcal{A}$.

- Modèle de Bernoulli : M0
  Les X_i sont indépendantes et générées avec les probabilités

$$\mu(a) = \mathbb{P}(X_i = a), \quad \forall a \in \mathcal{A}$$

  Peut s'ajuster sur la fréquence observée des lettres d'une séquence.

- Chaîne de Markov d'ordre 1 : M1
  Les X_i ne sont plus indépendantes mais générées selon

$$\begin{aligned} \mu(a) &= \mathbb{P}(X_1 = a), \quad \forall a \in \mathcal{A} \\ \pi(a, b) &= \mathbb{P}(X_i = b \mid X_{i-1} = a), \quad \forall a, b \in \mathcal{A} \end{aligned}$$

  Peut s'ajuster sur la fréquence observée des 2-mots d'une séquence.

Modèles de séquences classiques (2)

- Chaîne de Markov d'ordre m : Mm
  Les X_i dépendent des m lettres précédentes et sont générées selon

$$
\begin{aligned}
\mu(a_1 \cdots a_m) &= \mathbb{P}(X_1 \cdots X_m = a_1 \cdots a_m), \quad \forall a_j \in \mathcal{A} \\
\pi(a_1 \cdots a_m, b) &= \mathbb{P}(X_i = b \,|\, X_{i-1} \cdots X_{i-1} = a_1 \cdots a_m),
\end{aligned}
$$

  Peut s'ajuster sur la fréquence observée des $(m+1)$-mots d'une séquence.

Modèles de séquences classiques (2)

- Chaîne de Markov d'ordre m : Mm
  Les X_i dépendent des m lettres précédentes et sont générées selon

$$\begin{aligned} \mu(a_1 \cdots a_m) &= \mathbb{P}(X_1 \cdots X_m = a_1 \cdots a_m), \quad \forall a_j \in \mathcal{A} \\ \pi(a_1 \cdots a_m, b) &= \mathbb{P}(X_i = b \,|\, X_{i-1} \cdots X_{i-1} = a_1 \cdots a_m), \end{aligned}$$

  Peut s'ajuster sur la fréquence observée des $(m+1)$-mots d'une séquence.

Ces modèles sont basés sur une **hypothèse d'homogénéité** de la séquence : les probabilités d'émission des lettres sont identiques tout du long de la séquence.

Assouplir l'hypothèse d'homogénéité

- **Modèles markoviens avec phase** pour les séquences codantes : par exemple, $\pi(a, b)$ devient $\pi_1(a, b)$, $\pi_2(a, b)$ ou $\pi_3(a, b)$ selon que la lettre b est générée en position 1, 2 ou 3 d'un codon
$\Rightarrow$ M1_3.

Assouplir l'hypothèse d'homogénéité

- **Modèles markoviens avec phase** pour les séquences codantes : par exemple, $\pi(a,b)$ devient $\pi_1(a,b)$, $\pi_2(a,b)$ ou $\pi_3(a,b)$ selon que la lettre b est générée en position 1, 2 ou 3 d'un codon $\Rightarrow$ M1_3.
- **Chaînes de Markov hétérogènes** rarement utilisées pour une séquence (pb. d'estimation) :

$$\pi_i(a,b) = \mathbb{P}(X_i = b \mid X_{i-1} = a), \quad \forall a, b \in \mathcal{A}$$

Assouplir l'hypothèse d'homogénéité

- **Modèles markoviens avec phase** pour les séquences codantes : par exemple, $\pi(a, b)$ devient $\pi_1(a, b)$, $\pi_2(a, b)$ ou $\pi_3(a, b)$ selon que la lettre b est générée en position 1, 2 ou 3 d'un codon $\Rightarrow$ M1_3.
- **Chaînes de Markov hétérogènes** rarement utilisées pour une séquence (pb. d'estimation) :

$$\pi_i(a, b) = \mathbb{P}(X_i = b \mid X_{i-1} = a), \quad \forall a, b \in \mathcal{A}$$

- **Chaînes de Markov cachées** (CMC, *HMM en anglais*) : les probabilités d'émission à une position i dépendent de l'**état** de la position i (le nombre d'états est petit).
Exemples d'états : isochores, introns/exons/intergéniques, hélices/feuillets/boucles, etc.

Enjeux des CMC pour l’analyse des séquences

- **Si la succession des états est connue** : il s’agira simplement d’estimer les paramètres du modèle (probabilités d’émission). On pourra ensuite caractériser chacun des états ou utiliser ce modèle à des fins prédictives sur une autre séquence.

Enjeux des CMC pour l'analyse des séquences

- **Si la succession des états est connue** : il s'agira simplement d'estimer les paramètres du modèle (probabilités d'émission). On pourra ensuite caractériser chacun des états ou utiliser ce modèle à des fins prédictives sur une autre séquence.
- **Si la segmentation n'est pas connue** : il s'agira de déterminer celle qui correspond "au mieux" à la séquence observée, voire d'estimer les paramètres s'ils sont aussi inconnus.

Chaînes de Markov cachées

Présentation

Un modèle de CMC permet de modéliser une séquence par un ensemble fini de modèles qui s'alternent le long de la séquence



Présentation

Un modèle de CMC permet de modéliser une séquence par un ensemble fini de modèles qui s'alternent le long de la séquence



Il y a donc deux processus sous-jacents :

- Le processus non observable (caché) $S_1 S_2 S_3 \cdots S_n$ qui modélisera la suite des états le long de la séquence.

  Ce processus est une chaîne de Markov d'ordre 1.
  $\Rightarrow$ "Chaîne de Markov Cachée".

Présentation

Un modèle de CMC permet de modéliser une séquence par un ensemble fini de modèles qui s'alternent le long de la séquence



Il y a donc deux processus sous-jacents :

- Le processus non observable (caché) $S_1 S_2 S_3 \cdots S_n$ qui modélisera la suite des états le long de la séquence.

  Ce processus est une chaîne de Markov d'ordre 1.
  $\Rightarrow$ "Chaîne de Markov Cachée".

- Le processus observable $X_1 X_2 X_3 \cdots X_n$ qui modélisera la succession des lettres.

  Le modèle pour générer X_i dépend de l'etat S_i.

Présentation (2)

Dans le schéma ci-dessous, on peut distinguer trois états (**rouge**, **vert**, **bleu**) : les premières lettres suivent le modèle rouge, etc. les dernières le modèle vert.

X attaggcagatacgaggtgattactcgctagtct

S

Présentation (2)

Dans le schéma ci-dessous, on peut distinguer trois états (**rouge**, **vert**, **bleu**) : les premières lettres suivent le modèle rouge, etc. les dernières le modèle vert.

X attaggcagatac ga ggt gattactcgctagtct
S

Les régions rouges, vertes et bleues sont caractérisées par des lois d'apparition des bases différentes (par ex. les régions rouges sont riches en **g**, etc.).

L'alternance des couleurs (états) est régie par une chaîne de Markov d'ordre 1.

Chaîne de Markov d'ordre 1 : M1

Une chaîne de Markov est une suite de variables aléatoires **dépendantes**

$$S_1 S_2 S_3 \cdots S_n \cdots$$

Ici S_i peut prendre un nombre fini de valeurs $\mathcal{S}$ (par ex. $\{\mathbf{r}, \mathbf{v}, \mathbf{b}\}$).

Chaîne de Markov d'ordre 1 : M1

Une chaîne de Markov est une suite de variables aléatoires **dépendantes**

$$S_1 S_2 S_3 \cdots S_n \cdots$$

Ici S_i peut prendre un nombre fini de valeurs $\mathcal{S}$ (par ex. $\{\mathbf{r}, \mathbf{v}, \mathbf{b}\}$).

Une dépendance d'**ordre** 1 signifie :

$$\mathbb{P}(S_i = b \mid S_1, S_2, \ldots, S_{i-1}) = \mathbb{P}(S_i = b \mid S_{i-1}) \ ;$$

la valeur de S_{i-1} suffit pour connaître avec quelle probabilité S_i prend la valeur b.

Modèle M1 : Matrice de transition

Les S_i sont donc générées successivement selon les probabilités de transition :

$$\pi(u,v) = \mathbb{P}(S_i = v \mid S_{i-1} = u) \ ;$$

celles-ci sont rangées dans une matrice de transition Π.

Par exemple, $\mathcal{S} = \{\mathbf{r}, \mathbf{v}, \mathbf{b}\}$ et

$$\Pi = \begin{pmatrix} \mathbf{0.6} & 0.4 & 0 \\ 0.5 & \mathbf{0} & 0.5 \\ 0.3 & 0.5 & \mathbf{0.2} \end{pmatrix}$$

$$\begin{aligned} \mathbb{P}(S_i = \mathbf{v} \mid S_{i-1} = \mathbf{v}) &= 0 \\ \mathbb{P}(S_i = \mathbf{v} \mid S_{i-1} = \mathbf{r}) &= 0.4 \qquad \text{etc.} \end{aligned}$$

Propriété : les sommes en ligne de la matrice de transition font 1.

Modèle M1 : loi initiale

Pour démarrer la chaîne, il faut se donner une loi de probabilité pour la première couleur appelée **loi initiale** :

$$\mathbb{P}(S_1 = u), \quad \forall u \in \mathcal{S}$$

Une chaîne de Markov d'ordre 1 est donc définie par une loi initiale et une matrice de transition.

Modèle M1 : loi initiale

Pour démarrer la chaîne, il faut se donner une loi de probabilité pour la première couleur appelée **loi initiale** :

$$\mathbb{P}(S_1 = u), \quad \forall u \in \mathcal{S}$$

Une chaîne de Markov d'ordre 1 est donc définie par une loi initiale et une matrice de transition.

En pratique, la loi initiale est choisie comme étant la **loi stationnaire** $\mu(\cdot)$, c'est-à-dire vérifiant $\mu = \mu\Pi$.
Ceci garantit que les variables S_i ont la même loi μ :

$$\mathbb{P}(S_i = u) = \mu(u), \quad \forall i, \quad \forall u \in \mathcal{S}.$$

La chaîne est alors dite stationnaire.

Modèle M1 : estimation de Π

Comment choisir les probabilités de transition $\pi(u, v)$ si l'on dispose d'une suite de couleurs observée $s_1 s_2 \cdots s_n$?

Modèle M1 : estimation de Π

Comment choisir les probabilités de transition $\pi(u,v)$ si l'on dispose d'une suite de couleurs observée $s_1 s_2 \cdots s_n$?

L'**estimation par maximum de vraisemblance** consiste à choisir les paramètres $\pi(u,v)$ qui maximise la vraisemblance

$$\begin{aligned}\mathbb{P}(S_1 S_2 \cdots S_n = s_1 s_2 \cdots s_n \mid \Pi) \\ &= \mu(s_1)\pi(s_1, s_2) \times \cdots \times \pi(s_{n-1}, s_n) \\ &= \mu(s_1) \prod_{u,v \in \mathcal{S}} \big(\pi(u,v)\big)^{N_{\text{obs}}(uv)}.\end{aligned}$$

où $N_{\text{obs}}(uv)$ est le nombre d'occurrences de "uv" dans $s_1 s_2 \cdots s_n$

Modèle M1 : estimation de Π

Comment choisir les probabilités de transition $\pi(u,v)$ si l'on dispose d'une suite de couleurs observée $s_1 s_2 \cdots s_n$?

L'**estimation par maximum de vraisemblance** consiste à choisir les paramètres $\pi(u,v)$ qui maximise la vraisemblance

$$\begin{aligned}\mathbb{P}(S_1 S_2 \cdots S_n &= s_1 s_2 \cdots s_n \mid \Pi) \\ &= \mu(s_1)\pi(s_1,s_2) \times \cdots \times \pi(s_{n-1},s_n) \\ &= \mu(s_1) \prod_{u,v \in \mathcal{S}} \big(\pi(u,v)\big)^{N_{\text{obs}}(uv)}.\end{aligned}$$

où $N_{\text{obs}}(uv)$ est le nombre d'occurrences de "uv" dans $s_1 s_2 \cdots s_n$

$$\Longrightarrow \quad \widehat{\pi}(u,v) = \frac{N_{\text{obs}}(uv)}{N_{\text{obs}}(u+)}$$

Modèle M1 : ajustement sur les “di”

Dans des chaînes de Markov d’ordre 1 telles que $\pi(u,v) = \frac{N_{\text{obs}}(uv)}{N_{\text{obs}}(u+)}$, on a **en moyenne**

$$N(uv) \simeq N_{\text{obs}}(uv).$$

Modèle M1 : ajustement sur les “di”

Dans des chaînes de Markov d’ordre 1 telles que $\pi(u,v) = \frac{N_{\text{obs}}(uv)}{N_{\text{obs}}(u+)}$, on a **en moyenne**

$$N(uv) \simeq N_{\text{obs}}(uv).$$

De façon générale, le modèle Mm d’ordre m s’ajuste sur la composition en “mots” de taille 1, 2, . . ., $(m+1)$.

Paramètres d'un modèle CMC

- **Processus caché** $S = (S_1, S_2, S_3, \ldots, S_n)$**,** $S_i \in \mathcal{S}$ **(M1)**

$$
\begin{aligned}
\mu_e(u) &= \mathbb{P}(S_1 = u), \quad \forall u \in \mathcal{S} \\
\pi_e(u, v) &= \mathbb{P}(S_i = v \mid S_{i-1} = u), \quad \forall u, v \in \mathcal{S}
\end{aligned}
$$

Paramètres d'un modèle CMC

- **Processus caché** $S = (S_1, S_2, S_3, \ldots, S_n)$**,** $S_i \in \mathcal{S}$ **(M1)**

$$\begin{aligned} \mu_e(u) &= \mathbb{P}(S_1 = u), \quad \forall u \in \mathcal{S} \\ \pi_e(u,v) &= \mathbb{P}(S_i = v \mid S_{i-1} = u), \quad \forall u, v \in \mathcal{S} \end{aligned}$$

- **Processus observé** $X = (X_1, X_2, X_3, \cdots, X_n)$**,** $X_i \in \mathcal{A}$**.** Conditionnellement à S, le processus X peut suivre le modèle M0 ou M1 ou Mm ou un autre. Les modèles peuvent être de différentes natures selon les états.

Paramètres d'un modèle CMC

- **Processus caché** $S = (S_1, S_2, S_3, \ldots, S_n)$**,** $S_i \in \mathcal{S}$ **(M1)**

$$\begin{aligned} \mu_e(u) &= \mathbb{P}(S_1 = u), \quad \forall u \in \mathcal{S} \\ \pi_e(u, v) &= \mathbb{P}(S_i = v \mid S_{i-1} = u), \quad \forall u, v \in \mathcal{S} \end{aligned}$$

- **Processus observé** $X = (X_1, X_2, X_3, \cdots, X_n)$**,** $X_i \in \mathcal{A}$**.** Conditionnellement à S, le processus X peut suivre le modèle M0 ou M1 ou Mm ou un autre. Les modèles peuvent être de différentes natures selon les états.

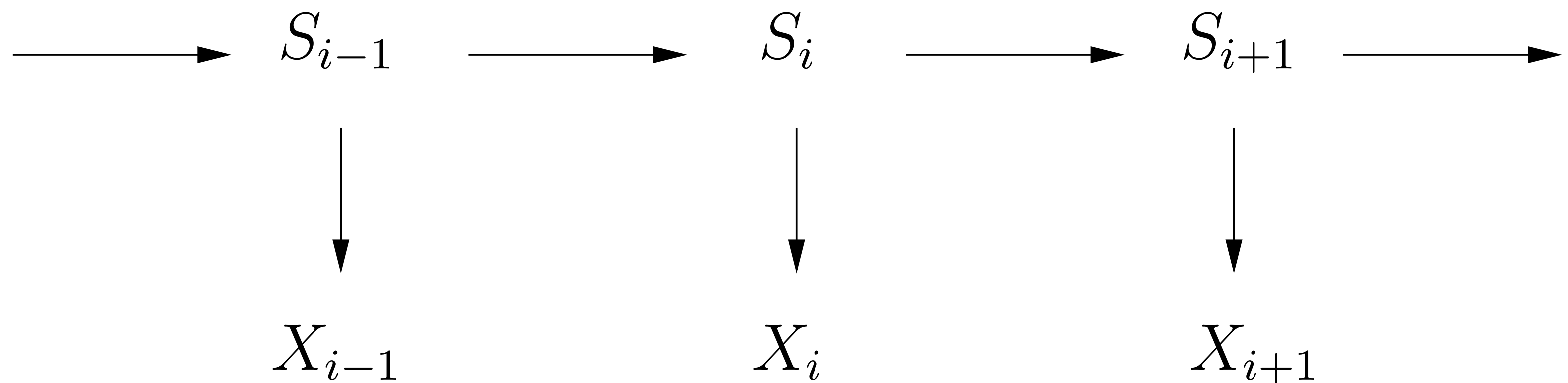
On se placera ici dans le cas où, condit. à S, X est markovien d'ordre $m \geq 0$ dans tous les états et on notera le modèle global **M1-M**m. Chaque état se caractérise donc par une certaine composition en oligos de taille 1 à $(m+1)$.

Exemple du modèle M1-M0

- **Paramètres pour les X_i** : les X_i sont indépendants, conditionnellement à S, et générés selon une loi μ_o

$$\mu_o(u, a) = \mathbb{P}(X_i = a \mid S_i = u), \ a \in \mathcal{A}, \ u \in \mathcal{S}$$

- **Schéma de dépendances** :



- **Simulation** : on simule d'abord $S = (S_1, S_2, S_3, \ldots, S_n)$ selon une CM d'ordre 1, puis on simule les X_i indépendamment des autres : X_i suit la loi d'émission de l'état s_i.

Exemple du modèle M1-M1

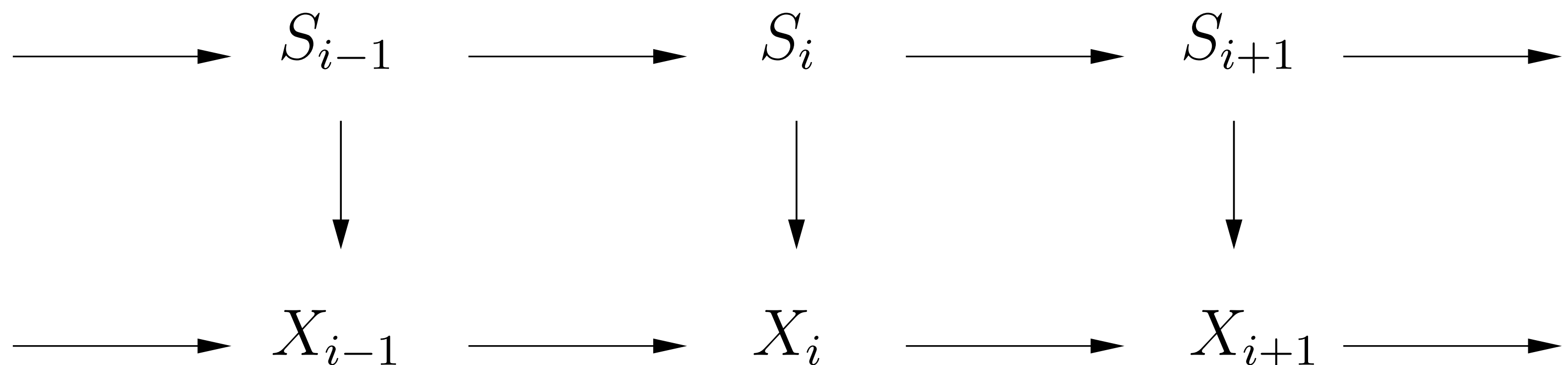
- **Paramètres pour les X_i** :
  les X_i forment une chaîne de Markov stationnaire d'ordre 1, conditionnellement à S, de loi initiale μ_o

$$\mu_o(u,a) = \mathbb{P}(X_1 = a \mid S_1 = u),\ a \in \mathcal{A},\ u \in \mathcal{S}$$

  et de matrice de transition

$$\pi_o(u,a,b) = \mathbb{P}(X_i = b \mid X_{i-1} = a, S_1 = u),\ a,b \in \mathcal{A},\ u \in \mathcal{S}$$

- **Schéma de dépendances** :



Représentation d'un modèle de CMC

On représente classiquement l'architecture d'un modèle de CMC par un graphe dont les noeuds désignent les états et les arêtes les transitions autorisées entre états.

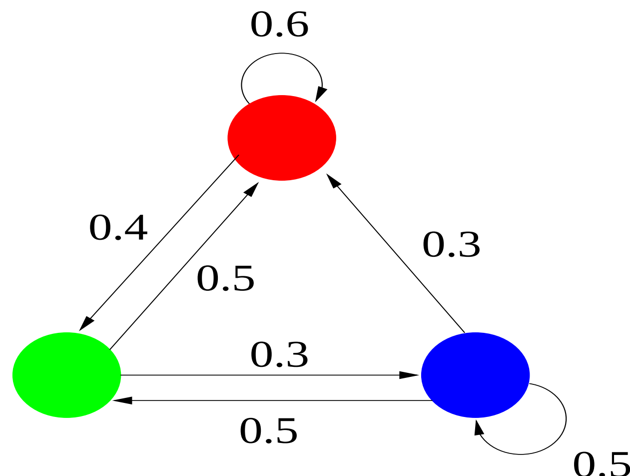
Représentation d'un modèle de CMC

On représente classiquement l'architecture d'un modèle de CMC par un graphe dont les noeuds désignent les états et les arêtes les transitions autorisées entre états.

Par exemple, le graphe associé au modèle M1-M0 suivant

$$\Pi_e = \begin{pmatrix} \mathbf{0.6} & 0.4 & 0 \\ 0.5 & \mathbf{0} & 0.5 \\ 0.3 & 0.5 & \mathbf{0.2} \end{pmatrix} \quad \mu_o = \begin{pmatrix} \mathbf{0.2} & \mathbf{0.2} & \mathbf{0.3} & \mathbf{0.3} \\ \mathbf{0.1} & \mathbf{0.3} & \mathbf{0.4} & \mathbf{0.2} \\ \mathbf{0.3} & \mathbf{0.2} & \mathbf{0.4} & \mathbf{0.1} \end{pmatrix}$$

est



Estimation / Segmentation

Il y a deux écoles.

- L'approche **supervisée** consiste
  - à estimer les paramètres du modèle sur des séquences déjà segmentées
    $=$ maximum de vraisemblance sachant (X, S).
    $\longrightarrow$ implique de connaître la signification des états.
  - puis à segmenter la séquence d'intérêt avec ces paramètres.
    $=$ algorithme de Viterbi, ou algorithme "forward-backward".
- L'approche **non supervisée** consiste à itérer les deux étapes d'estimation/segmentation directement sur la séquence
  $=$ algorithme EM.

Estimation à segmentation connue

“apprentissage supervisé”

Si la segmentation S est connue, alors la vraisemblance des données est simple et on peut la maximiser analytiquement :

Estimation à segmentation connue

“apprentissage supervisé”

Si la segmentation S est connue, alors la vraisemblance des données est simple et on peut la maximiser analytiquement :

$$\mathbb{P}(X \mid \theta, S) \overset{M1-M0}{=} \mu_e(S_1)\pi_e(S_1, S_2)\ldots\pi_e(S_{n-1}, S_n) \times \mu_o(S_1, X_1)\ldots\mu_o(S_n, X_n)$$

Estimation à segmentation connue

"apprentissage supervisé"

Si la segmentation S est connue, alors la vraisemblance des données est simple et on peut la maximiser analytiquement :

$$\begin{aligned}\mathbb{P}(X \mid \theta, S) &\overset{M1-M0}{=} \mu_e(S_1)\pi_e(S_1, S_2)\ldots\pi_e(S_{n-1}, S_n) \\ &\qquad \times \mu_o(S_1, X_1)\ldots\mu_o(S_n, X_n) \\ &= \mu_e(S_1) \prod_{u,v=1}^{q} \pi_e(u,v)^{N(uv)} \times \prod_{u=1}^{q} \prod_{a \in \mathcal{A}} \mu_o(u,a)^{N(u,a)}\end{aligned}$$

où $N(uv)$ est le nombre d'états u suivis de l'état v, et $N(u,a)$ est le nombre de lettres a dans l'état u.

Estimation à segmentation connue (2)

Pour maximiser la vraisemblance, on annule simultanément les dérivées partielles par rapport aux $\pi_e(u,v)$ et $\mu_o(u,a)$ et on obtient les estimateurs naturels :

$$\widehat{\pi}_e(u,v) = \frac{N(uv)}{N(u)}$$

$$\widehat{\mu}_o(u,a) = \frac{N(u,a)}{N(u)}$$

Segmentation à paramètres connus

C'est typiquement le cas quand on estime les paramètres sur un jeu de test déjà segmenté (cf. paragraphe précédent), et que l'on veut segmenter une nouvelle séquence en gardant ces valeurs de paramètres θ.

Segmentation à paramètres connus

C'est typiquement le cas quand on estime les paramètres sur un jeu de test déjà segmenté (cf. paragraphe précédent), et que l'on veut segmenter une nouvelle séquence en gardant ces valeurs de paramètres θ.

Etant donné $X = (X_1, X_2, X_3, \ldots, X_n)$ et θ, on cherche la suite d'états $(s_1^*, s_2^*, \ldots, s_n)$ la plus probable, c'est-à-dire celle qui maximise

$$\mathbb{P}(S_1 = s_1, \ldots, S_n = s_n \mid X, \theta)$$

ou encore (formule de Bayes)

$$\mathbb{P}(X_1, \ldots, X_n, S_1 = s_1, \ldots, S_n = s_n \mid \theta)$$

$\Rightarrow$ Algorithme de Viterbi.

Algorithme de Viterbi

Soit $\mathbb{P}^* = \max_{s_1,\ldots,s_n} \mathbb{P}(X_1,\ldots,X_n,S_1=s_1,\ldots,S_n=s_n \mid \theta)$.

$$\mathbb{P}^* = \max_{v} \underbrace{\max_{s_1,\ldots,s_{n-1}} \mathbb{P}(X_1,\ldots,X_n,S_1=s_1,\ldots,S_{n-1}=s_{n-1},S_n=v \mid \theta}_{Z_n(v)}$$

$$s_n^* = \arg\max_v Z_n(v)$$

Récurrence pour calculer $Z_i(v)$:

$$\begin{cases} Z_1(v) & = & \mathbb{P}(X_1, S_1 = v) = \mu_e(v)\mu_0(v, X_1) \\ Z_i(v) & = & \max_u \Big(Z_{i-1}(u)\pi_e(u,v) \Big) \mu_o(v, X_i) \end{cases}$$

$$s_{i-1}^* = \arg\max_u (Z_{i-1}(u)\pi_e(u, s_i^*))$$

Alternative : algorithme *Forward-Backward*

Plutôt que de calculer la suite d'états "optimale" $(s_1^*, s_2^*, \ldots, s_n)$, l'algorithme *Forward-Backward* permet de calculer les probabilités de se trouver dans l'état u à chaque position i, sachant (X, θ) :

$$\mathbb{P}(S_i = u \mid X = x, \theta), \quad i = 1, \ldots, n, \quad u \in \mathcal{S}.$$

Estimation à segmentation inconnue

"apprentissage non supervisé"

Si la segmentation S n'est pas connue, la vraisemblance $\mathbb{P}(X \mid \theta)$ n'est pas manipulable. Pour la maximiser, on utilise des *algorithmes itératifs* qui permettent d'approcher l'estimateur $\widehat{\theta}$ du maximum de vraisemblance.

Estimation à segmentation inconnue

"apprentissage non supervisé"

Si la segmentation S n'est pas connue, la vraisemblance $\mathbb{P}(X \mid \theta)$ n'est pas manipulable. Pour la maximiser, on utilise des *algorithmes itératifs* qui permettent d'approcher l'estimateur $\widehat{\theta}$ du maximum de vraisemblance.

L'algorithme EM (*Expectation-Maximization*) est le plus populaire. Clé : à chaque étape, la vraisemblance croît.

- point de départ $\theta^{(0)}$
- itération k alterne une étape E et une étape M
- critère d'arrêt :
  $|\log \mathbb{P}(X = x \mid \theta^{(k+1)}) - \log \mathbb{P}(X = x \mid \theta^{(k)})| < \varepsilon$ ou $k > M$

Estimation à segmentation inconnue

“apprentissage non supervisé”

Si la segmentation S n’est pas connue, la vraisemblance $\mathbb{P}(X \mid \theta)$ n’est pas manipulable. Pour la maximiser, on utilise des *algorithmes itératifs* qui permettent d’approcher l’estimateur $\widehat{\theta}$ du maximum de vraisemblance.

L’algorithme EM (*Expectation-Maximization*) est le plus populaire. Clé : à chaque étape, la vraisemblance croît.

- point de départ $\theta^{(0)}$
- itération k alterne une étape E et une étape M
- critère d’arrêt :
  $|\log \mathbb{P}(X = x \mid \theta^{(k+1)}) - \log \mathbb{P}(X = x \mid \theta^{(k)})| < \varepsilon$ ou $k > M$

Attention : pb. des maxima locaux $\Rightarrow$ plusieurs points de départ.

Algorithme EM

Une itération de l'algorithme :

- Etape E : on calcule $\mathbb{P}(S_i = u \mid X = x, \theta^{(k)})$, $i = 1, \ldots, n$, $u \in \mathcal{S}$ (algorithme *Forward-Backward*)
- Etape M : on calcule $\theta^{(k+1)}$ en utilisant la segmentation obtenue

$$
\begin{aligned}
\pi_e^{(k+1)}(u,v) &= \frac{\sum_i \mathbb{P}(S_i = u, S_{i+1} = v \mid X = x, \theta^{(k)})}{\sum_i \mathbb{P}(S_i = u \mid X = x, \theta^{(k)})} \\
\mu_o^{(k+1)}(u,a) &= \frac{\sum_i \mathbb{1}\{X_i = a\}\mathbb{P}(S_i = u \mid X = x, \theta^{(k)})}{\sum_i \mathbb{P}(S_i = u \mid X = x, \theta^{(k)})}
\end{aligned}
$$

Application à l’analyse de séquences

Recherche de régions homogènes

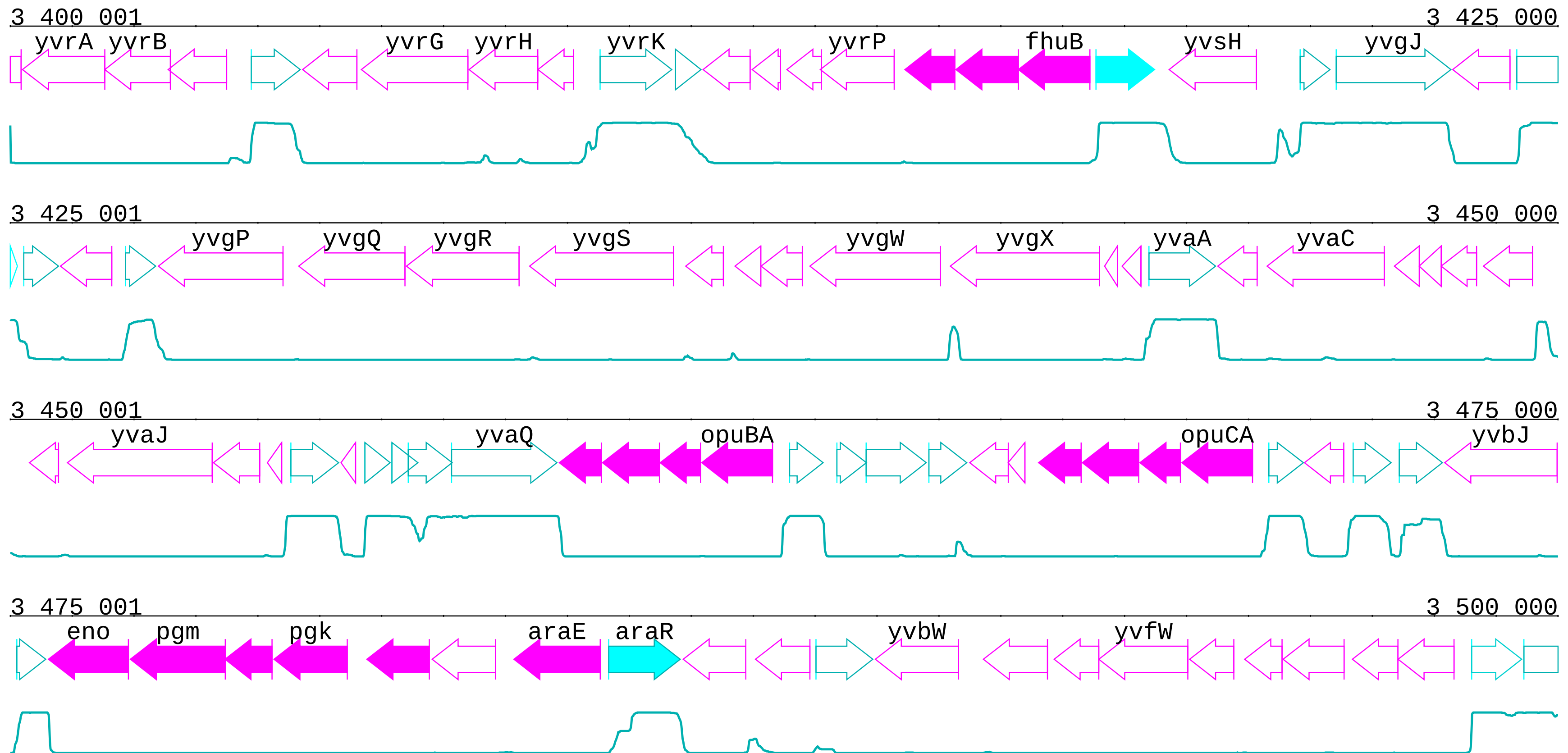
- On se donne un nombre q d'états cachés *a priori* (peu de résultats encore performants pour estimer ce nombre)
- On se donne les ordres des modèles markoviens sur les lettres, ou m.
- On applique EM pour estimer les paramètres et calculer les probabilités des états cachés en chaque site i de la séquence.

Recherche de régions homogènes (2)

(génomes bactériens)

Recherche de régions homogènes (3)

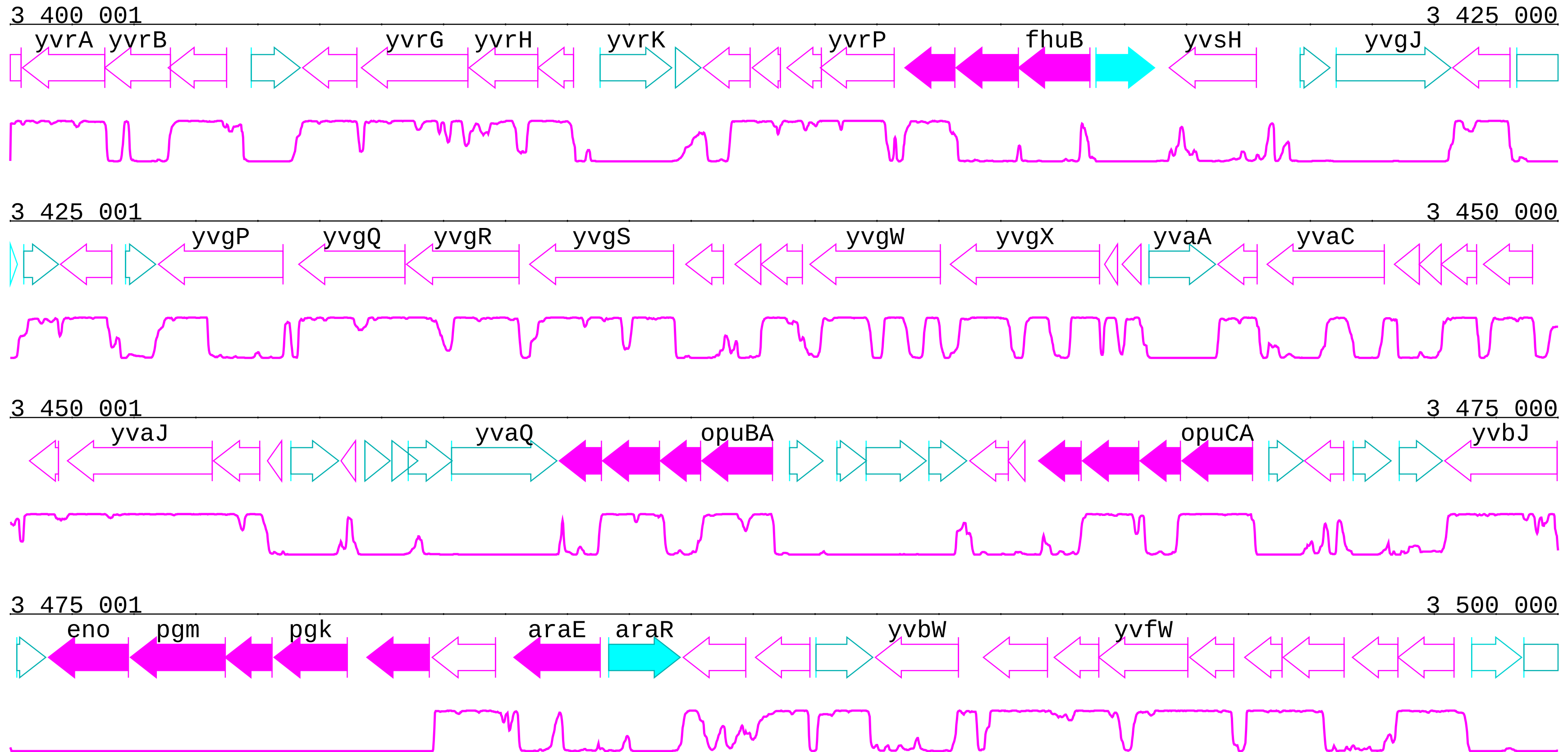
B. subtilis



état cyan : cds+

Recherche de régions homogènes (3)

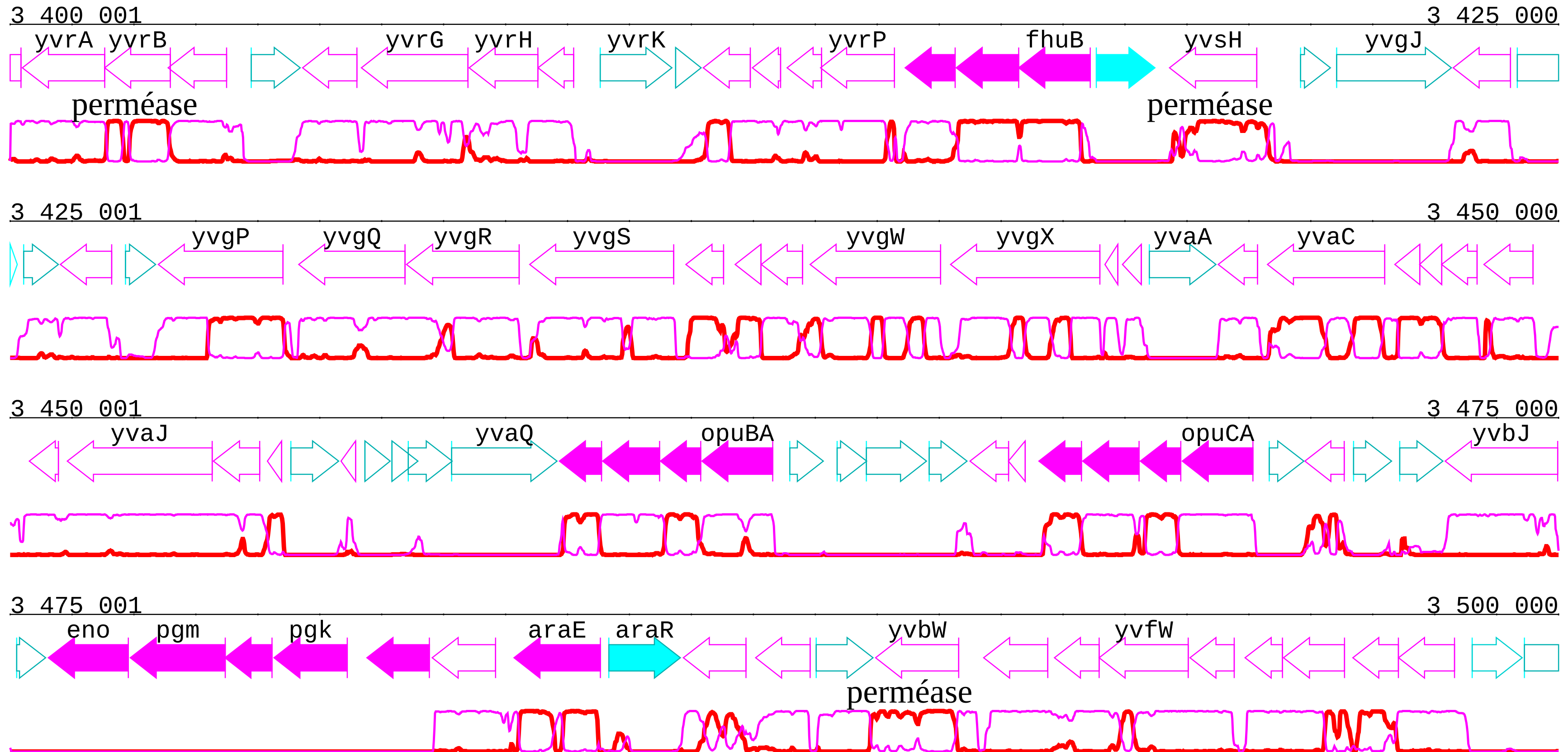
B. subtilis



état magenta : cds−

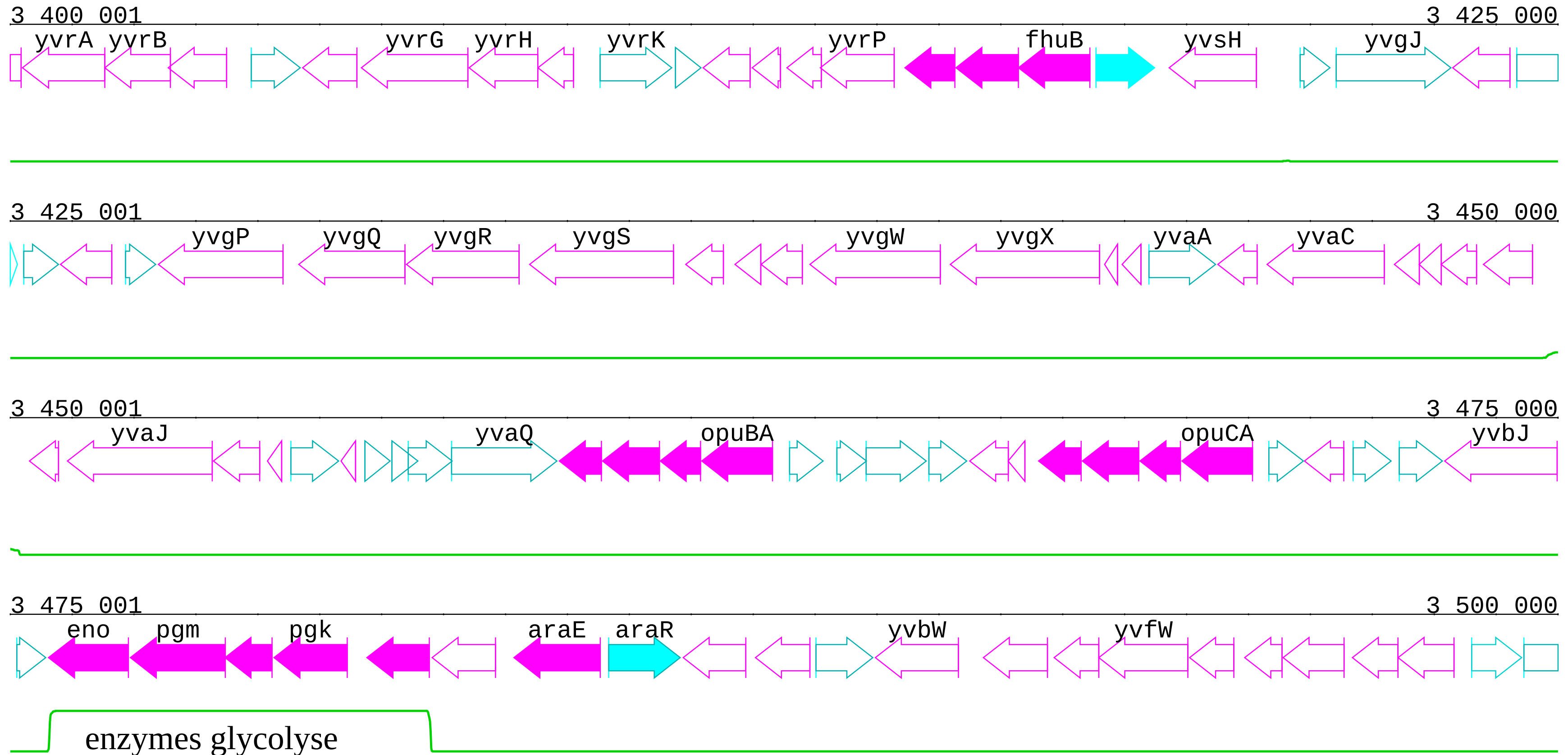
Recherche de régions homogènes (3)

B. subtilis



état rouge :cds− hydrophobe

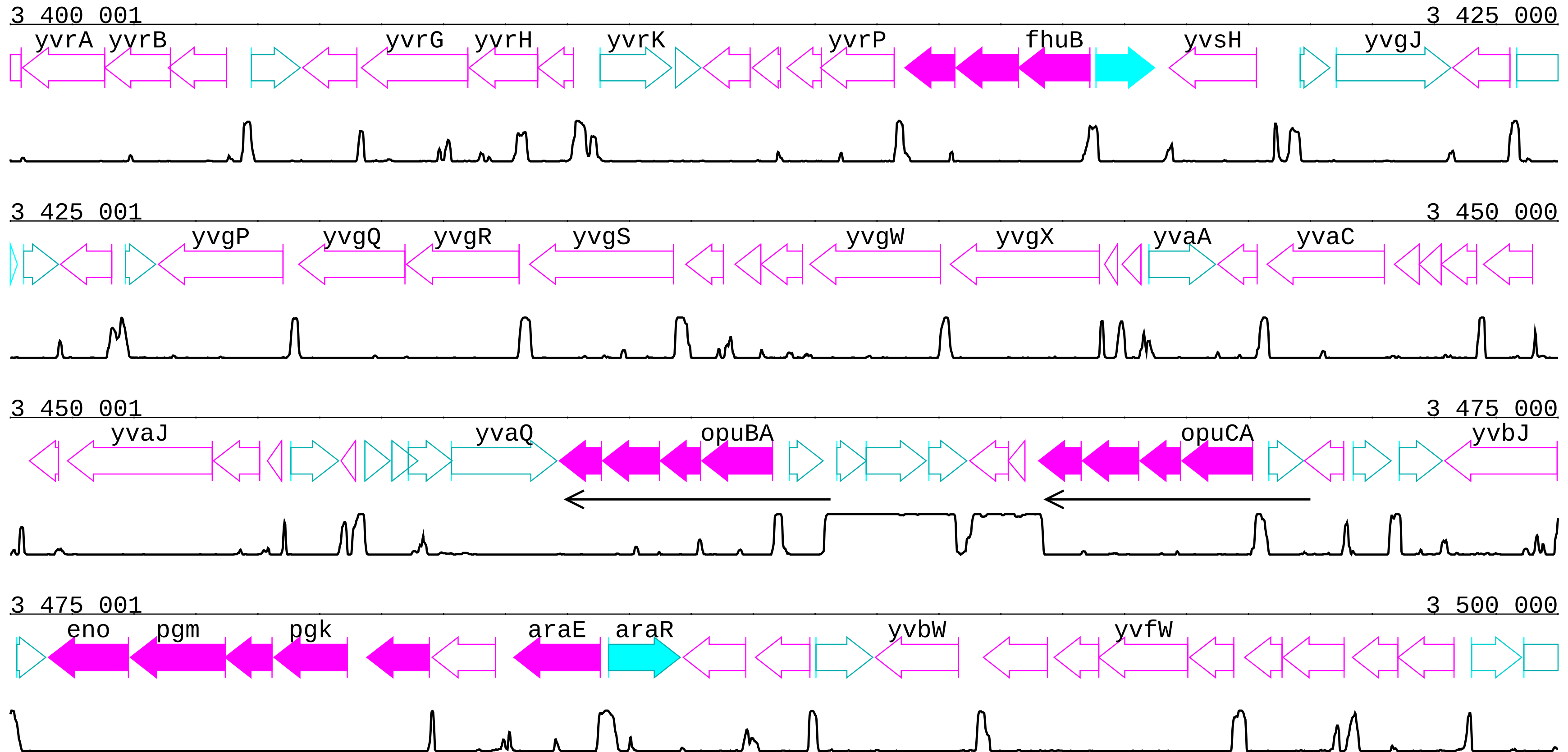
Recherche de régions homogènes (3)



état vert : fortement exprimé (enzyme de la glycolyse 4ième ligne)

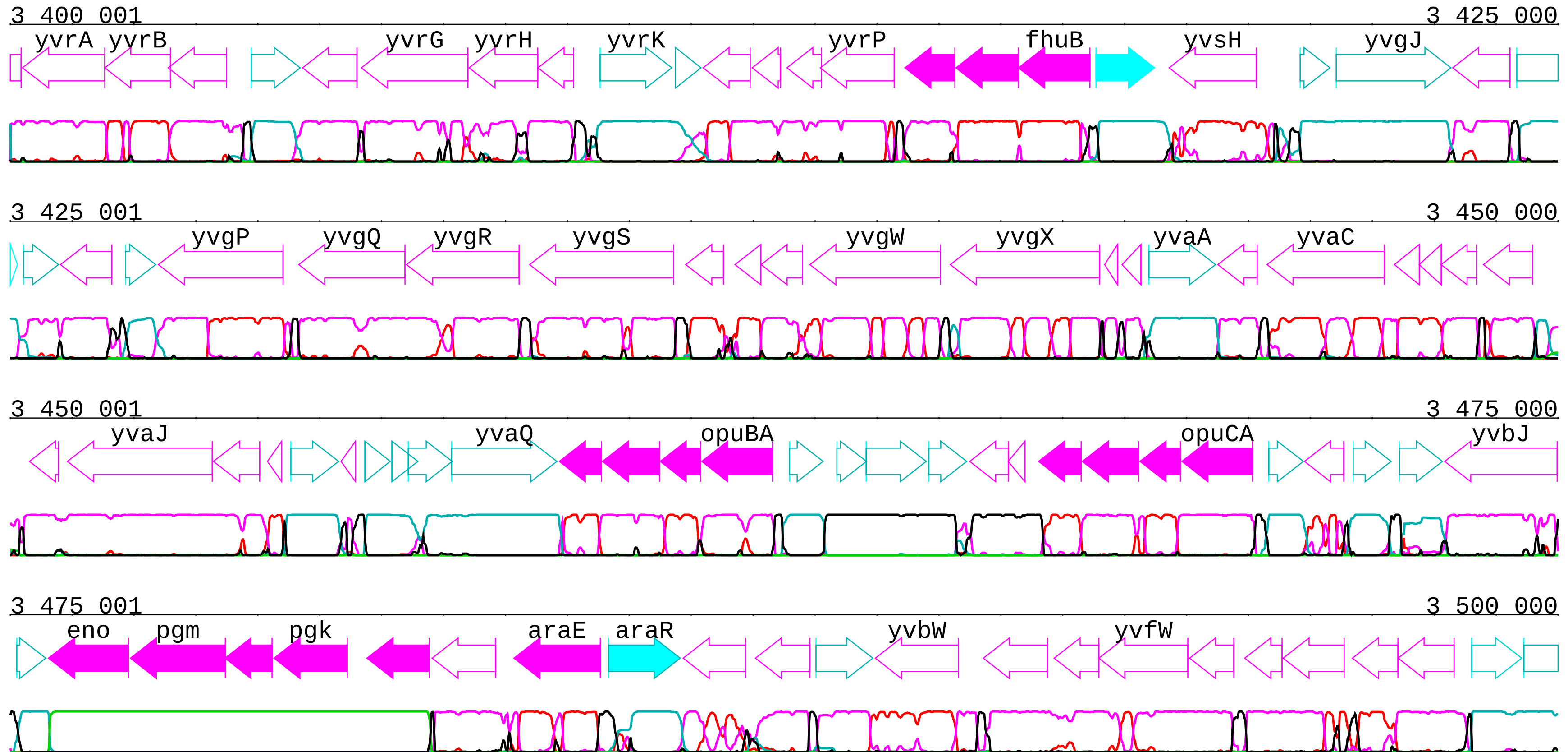
Recherche de régions homogènes (3)

B. subtilis



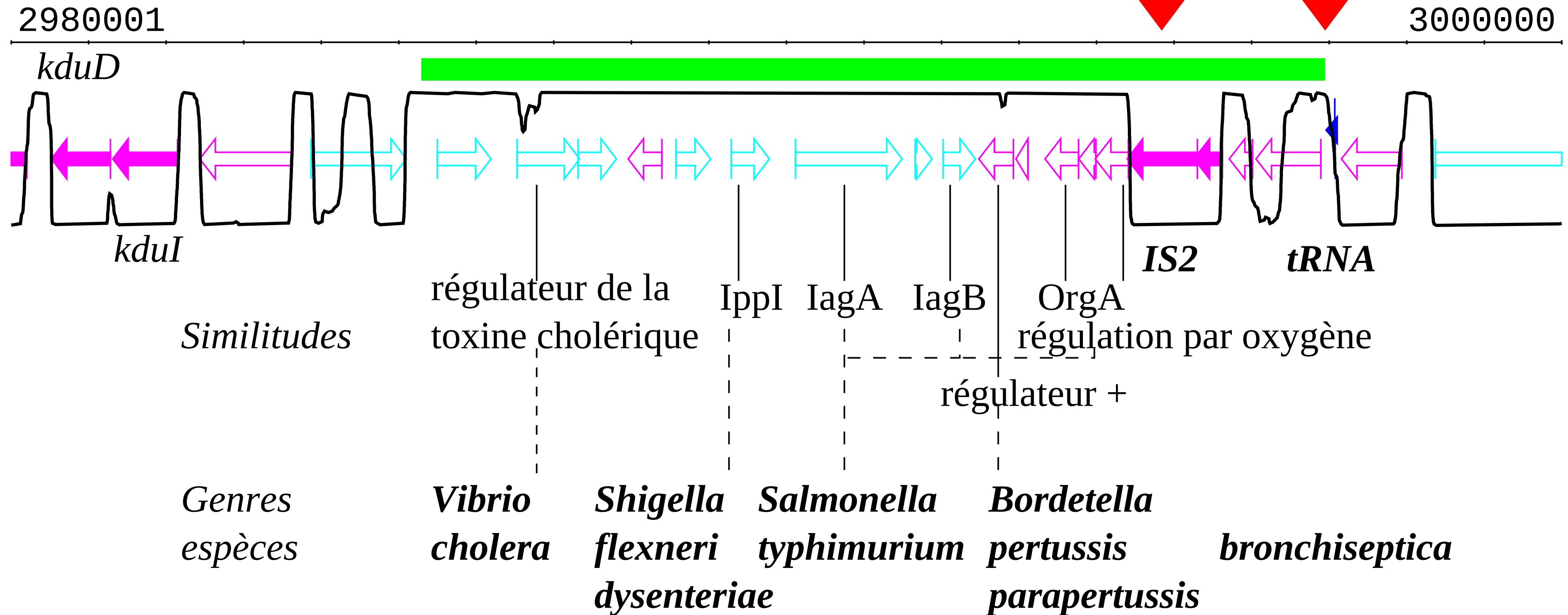
état noir : a+t-riche (transfert + intergenique)

Recherche de régions homogènes (3)



noir: a+t-riche (transfert + intergenique), vert: fortement exprime (enzyme de la glycolyse 4ieme ligne), cyan: cds+, magenta: cds-, rouge: cds- hydrophobe. La grande region en noir sur la 3ieme ligne est entouree d'une repetition (opBA=opuCA)

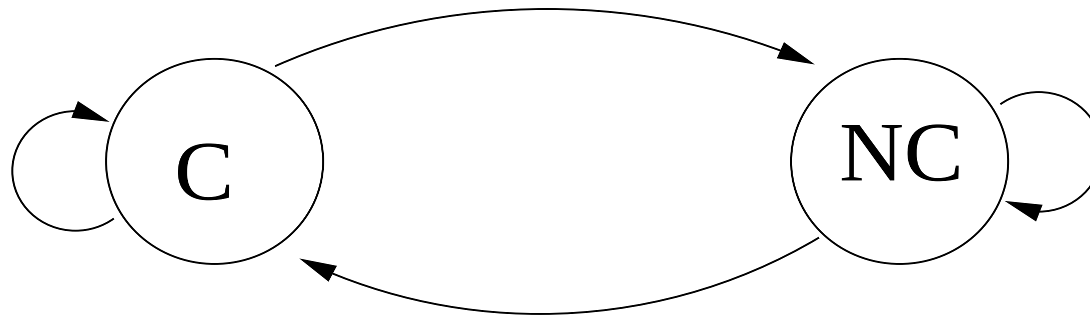
Recherche de régions homogènes (4)



Détection de gènes

De nombreux “détecteurs de gènes”.

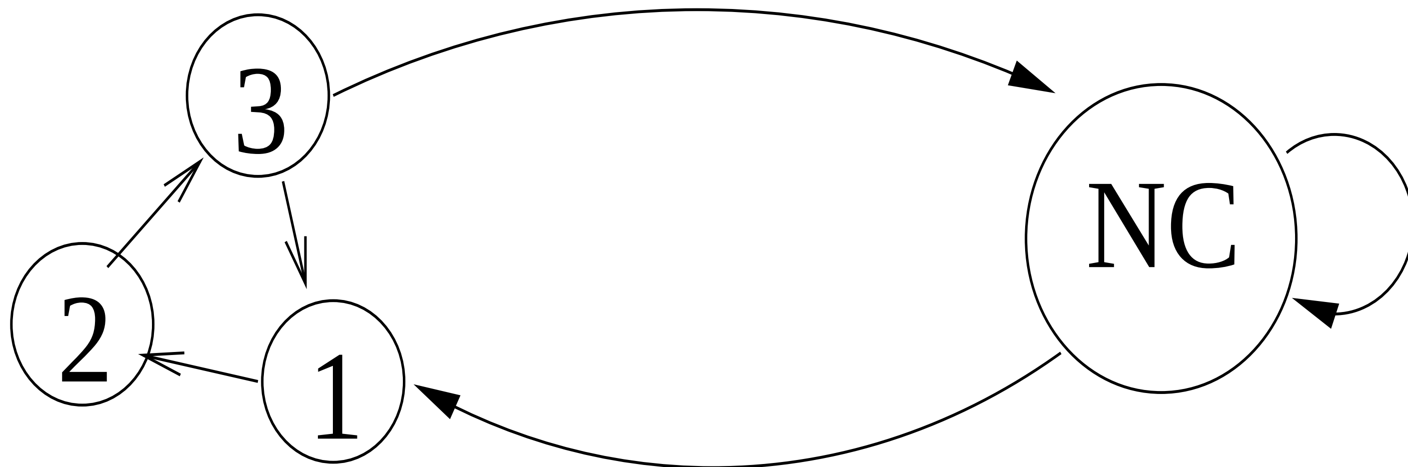
Principe de base : alternance de codant/intergénique (procaryotes) ou intergénique/exons/introns (eucaryotes), prise en compte de la phase pour le codant.



Détection de gènes

De nombreux “détecteurs de gènes”.

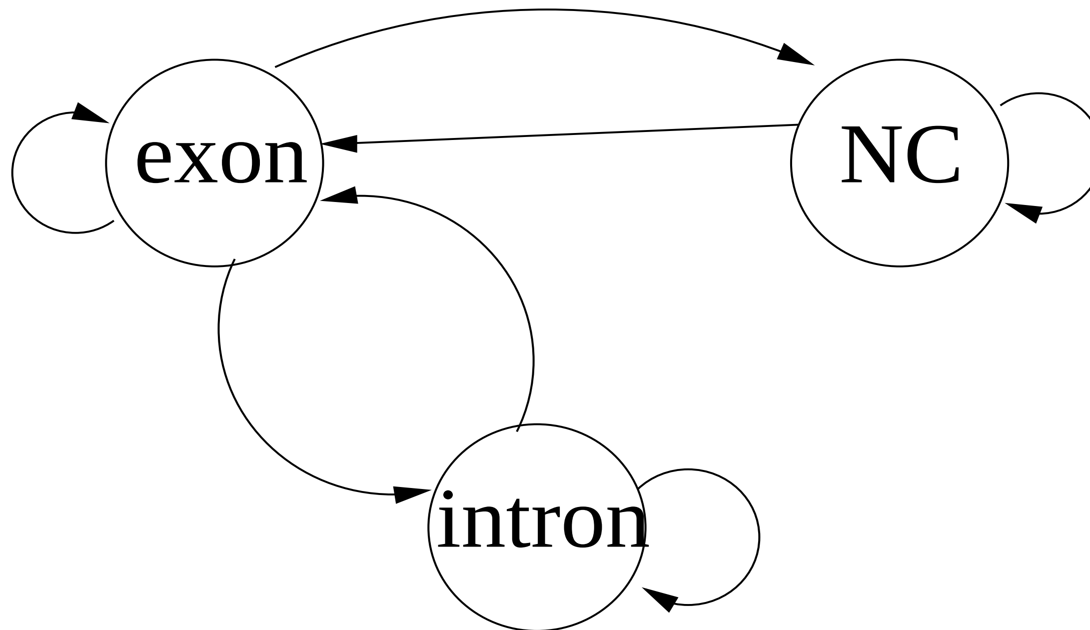
Principe de base : alternance de codant/intergénique (procaryotes) ou intergénique/exons/introns (eucaryotes), prise en compte de la phase pour le codant.



Détection de gènes

De nombreux “détecteurs de gènes”.

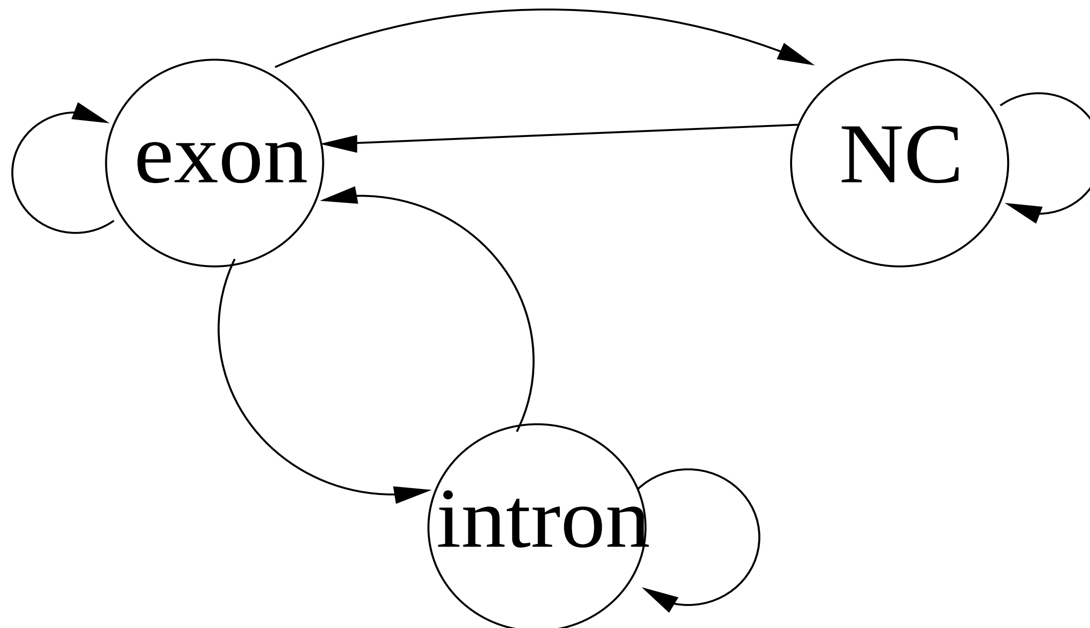
Principe de base : alternance de codant/intergénique (procaryotes) ou intergénique/exons/introns (eucaryotes), prise en compte de la phase pour le codant.



Détection de gènes

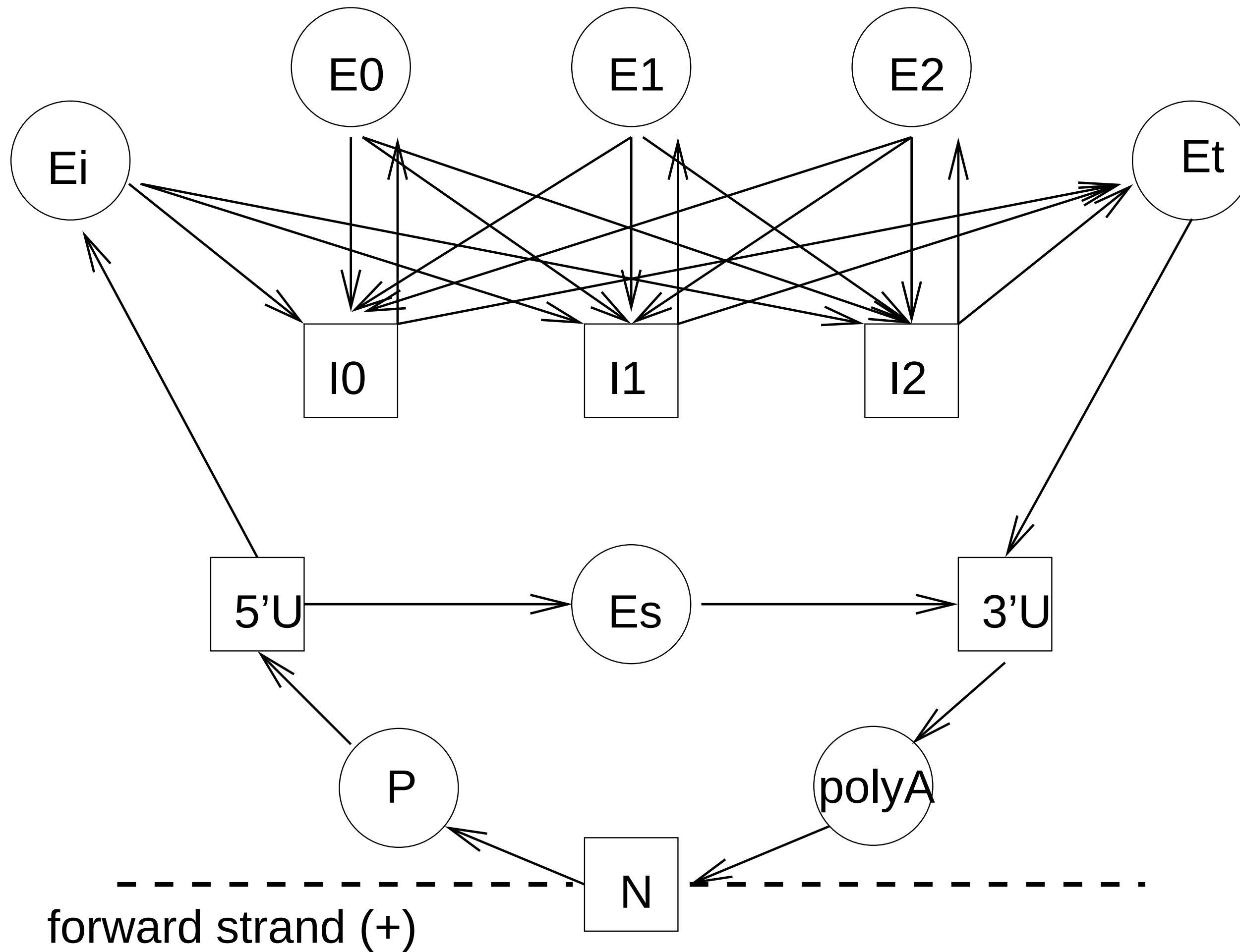
De nombreux “détecteurs de gènes”.

Principe de base : alternance de codant/intergénique (procaryotes) ou intergénique/exons/introns (eucaryotes), prise en compte de la phase pour le codant.



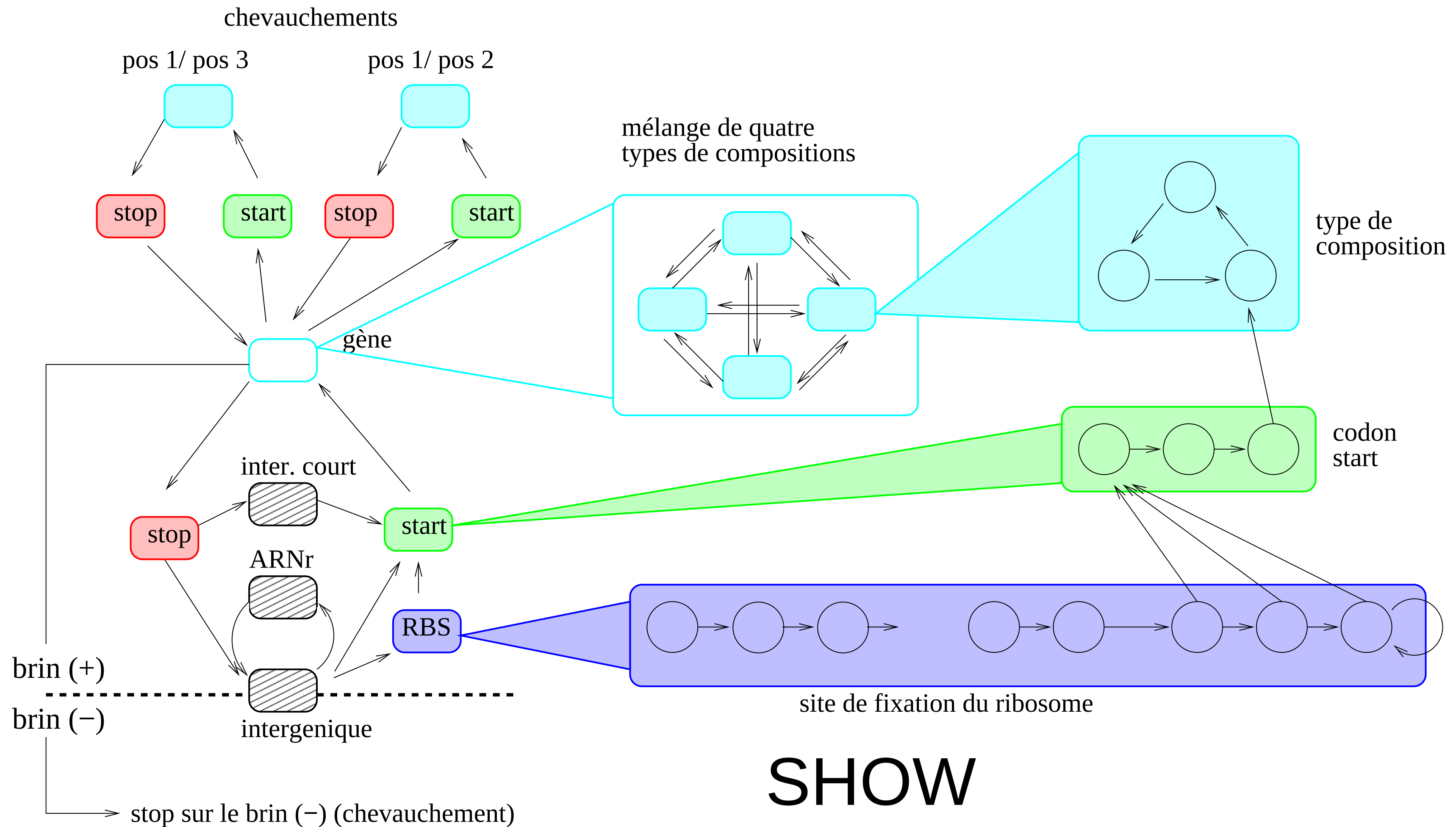
Complexifications : codons start/stop, exons initial/centraux/terminal, etc.

Détection de gènes (2)



(Genscan)

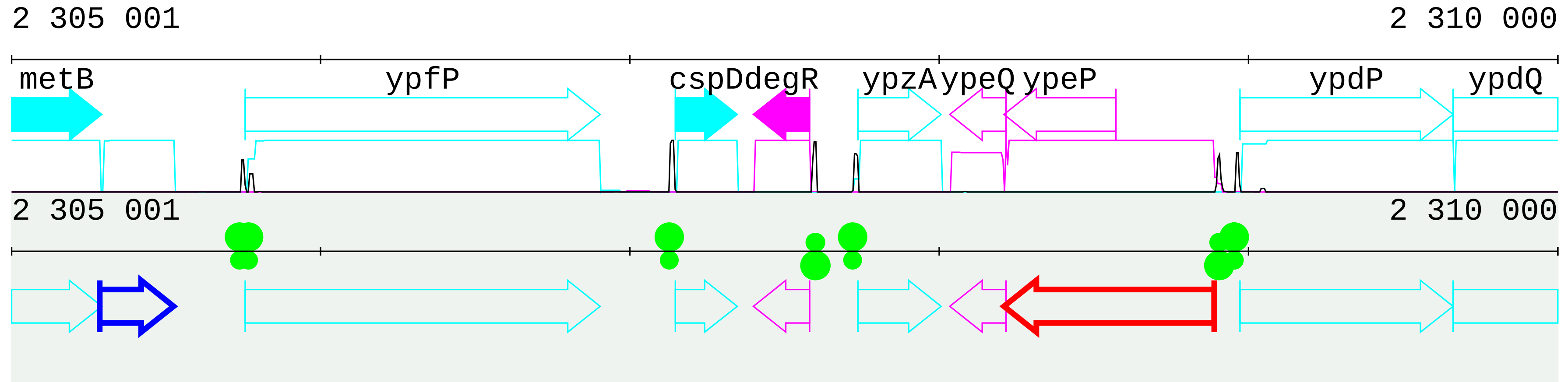
Détection de gènes (2)

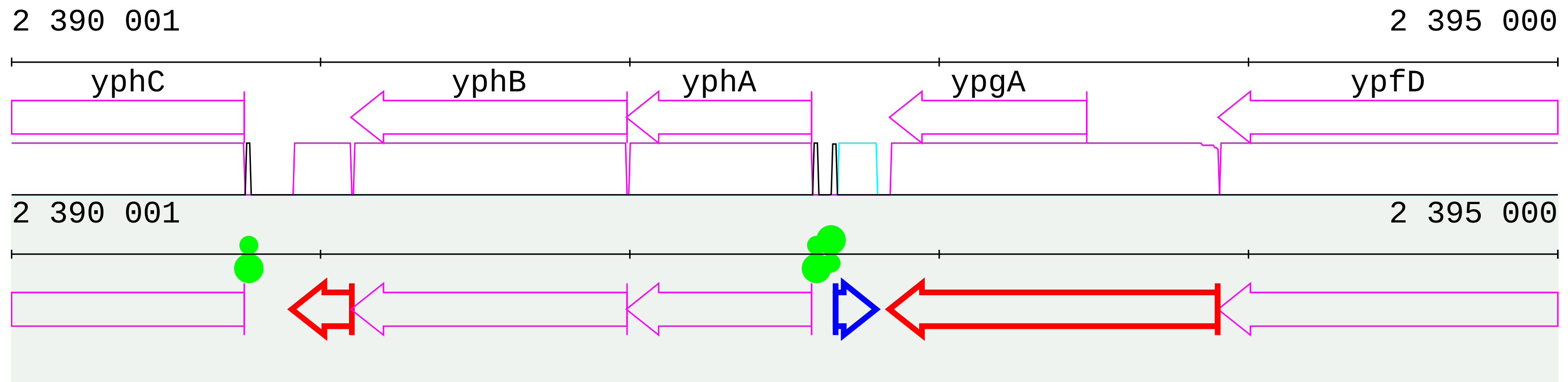


SHOW

Détection de gènes (3)

Petits gènes et départs de traduction





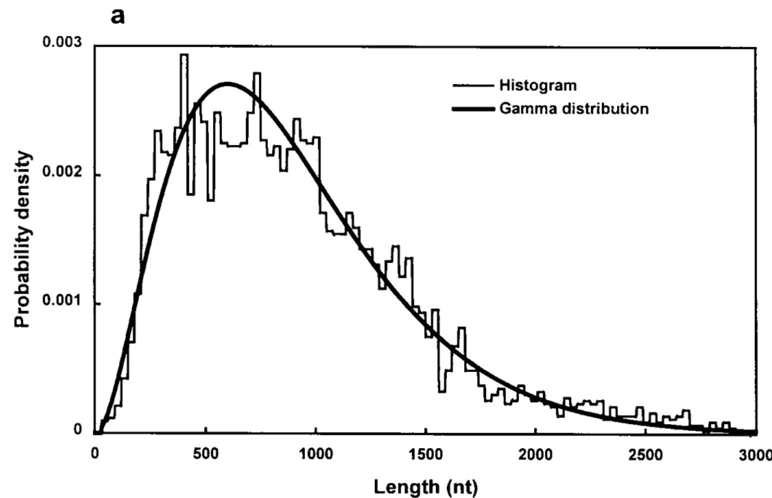
Modèles semi-Markov cachés

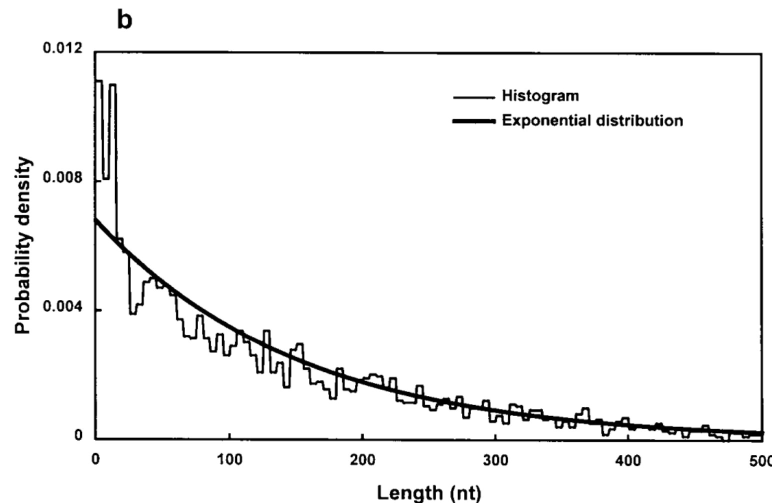
- L'inconvénient dans une CMC est que le temps de séjour est nécessairement distribué selon une loi géométrique :

$$\mathbb{P}(\text{temps } t \text{ dans l'état } u) = \Big(\pi_e(u,u)\Big)^{t-1}\big(1-\pi_e(u,u)\big).$$

- Les modèles semi-markovien (caché) permettent d'imposer une loi particulière (et adaptée au pb. biologique) pour le temps de séjour dans chaque état.

Exemple : longueurs d'exons/introns





(GeneMark.hmm, Genscan)